Cogn-IQ Research Papers Methodological Framework ◆ Open Access

doi: pubscience.org/ps-1n0JE-2ebae9-MqDI

THE DIFFICULTY HORSESHOE AS LOCAL INFORMATION

The Difficulty Horseshoe in Item-Level Multidimensional Scaling as a Signature of Local Information

Xavier Jouve[1]

[1]Cogn-IQ

Author Note

Xavier Jouve, Head of Research, Cogn-IQ. ORCID iD: 0009-0006-1283-045X.

Xavier Jouve is the developer of the Cogn-IQ instruments analyzed here (the EPC and the General Information Exam); this potential conflict of interest is disclosed for transparency. WAIS Information is a product of NCS Pearson, Inc. No external funding was received; no item content is reproduced.

Correspondence concerning this article should be addressed to Xavier Jouve, Cogn-IQ. Email: research@cogn-iq.org

ABSTRACT

When test items are scaled from their pairwise proximities, the configuration often bends into the classic difficulty horseshoe: an open arc running from the easiest items to the hardest. We show, on two operational item pools and a calibrated simulation, that this fold is best explained by the proximity coefficient's *information horizon* — how far along the difficulty continuum it remains informative before saturating — rather than by a property of the item data. On a 37-item pool ($N = 1{,}818$), phi proximities scaled at the ordinal level produce the textbook horseshoe; simple-matching proximities, which remain informative across the whole span, lay the same continuum out almost flat while fitting better. On a wide 178-item pool ($N = 483$), phi fails, and no correction within the phi family survives: margin-corrected and margin-free coefficients delete the difficulty dimension, and trimming extreme items shortens it. A Rasch simulation from the pool's own calibration shows phi requires samples in the thousands, whereas simple matching is indifferent to sample size. Restricting the scaling to local proximities (saturated long-range values treated as missing, as in archaeological seriation) reproduces the horseshoe on both pools under an explicit cutoff. A permutation control (marginals preserved, associations destroyed) fixes the interpretation: agreement-based configurations order items by difficulty from the marginals alone, whereas phi requires real

association structure; neither the ordering nor the fold is, by itself, evidence of a latent continuum. Across the pools studied here, the phi horseshoe is a consequence of coefficient saturation; an equivalent display can be produced explicitly, with the locality threshold stated and dimensionality established independently.

Keywords: multidimensional scaling, horseshoe effect, seriation, item difficulty, proximity coefficient

1. Introduction

Multidimensional scaling (MDS) of inter-item proximities is an established exhibit in psychometric documentation. When the items of a test are embedded in two dimensions, the configuration frequently takes the form of a curved band ordered by item difficulty — the horseshoe — and this shape is conventionally interpreted as evidence that a single continuum underlies the item set. The same geometry arises in other fields whenever objects ordered along one dimension are embedded in two, including archaeological seriation (Kendall, 1970) and spectral embeddings under local kernels (Diaconis, Goel, & Holmes, 2008).

The interpretation is complicated by a dependence that is rarely examined: the shape of the configuration varies with the proximity coefficient while the data are held constant. For a single set of 37 dichotomously scored items answered by 1,818 examinees, inter-item phi correlations scaled at the ordinal level produce the classic horseshoe; simple-matching proximities computed on the same responses produce a nearly linear configuration with better fit; and simple-matching proximities with the lowest quartile of values excluded from the fit reproduce the horseshoe at greater strength (Figure 2). The three analyses differ only in the definition of proximity and in the subset of proximities fitted.

Because the shape varies under the analyst's choices while the data are constant, the horseshoe cannot, by itself, be attributed to the item set. A companion paper gives a spectral account of the composition of the axes once a configuration exists — the first dimension as the difficulty fundamental, the arch as its first harmonic, content contrasts as separate irreducible components (Jouve, 2026b). The present paper addresses the prior

question of the conditions under which the fold exists, and whether it can be produced under explicit analytic control. We ask:

RQ1. Is the difficulty horseshoe a property of the item data, or of the proximity coefficient through which the data are viewed?
RQ2. Why does the classic phi-based recipe yield the horseshoe on some item pools and fail entirely on others — and can it be repaired within the phi family?
RQ3. Can the open horseshoe be produced under explicit analytic control, and what does such a configuration establish — and not establish — about the item set?

2. The Horseshoe in Theory

Nonmetric multidimensional scaling represents objects as points such that inter-point distances reproduce, up to a monotone transformation, the observed dissimilarities (Kruskal, 1964); the alternating least squares formulation used throughout this paper adds interval and ordinal measurement levels and a principled treatment of missing proximities (Takane, Young, & de Leeuw, 1977). When the objects differ along a single ordered dimension, their two-dimensional configuration is frequently curved. Kendall (1970) documented the phenomenon in archaeological seriation, where similarity data are informative only *locally*: assemblages share types only with assemblages of nearby date, and such data order the objects correctly while curving the ordering into an arc. Diaconis et al. (2008) formalized the mechanism for kernels that decay with distance: the embedding coordinates are eigenfunctions of a local averaging operator, the second eigenfunction is a quadratic bend of the first, and the embedded chain therefore arches.

The condition doing the work in both accounts is *locality*: the proximity must stop discriminating beyond some horizon. If a coefficient discriminates along the entire continuum — if objects twice as far apart reliably look twice as dissimilar — nothing forces a fold; a one-dimensional chain embeds happily along one dimension, and the second dimension is left to noise or to genuinely separate structure. The psychometric literature has long discussed which association coefficient to compute between dichotomous items, mostly in terms of factor-analytic artifacts: phi is bounded away from unity for items with unequal marginals, the classic "difficulty factor" problem, with the

bound given by Cureton's (1959) phi/phi-max and its later reappraisals (Davenport & El-Sanhurry, 1991). That literature did not address the geometric consequence for item configurations: the marginal bound makes phi a saturating, locally informative kernel along the difficulty continuum, which is the condition under which scaling curves a chain.

3. A Locality Account

Consider items scored 0/1, ordered by their proportions correct $p_1 \geq p_2 \geq \cdots \geq p_J$, and suppose for a moment that responding is perfectly cumulative in the scalogram sense of Guttman (1950): an examinee who passes an item passes every easier item. For two items i, j, the simple-matching similarity — the proportion of examinees giving the same response to both — is then an identity:

$$s_{ij}^{\mathrm{SM}} \;=\; P(\text{both } 1) + P(\text{both } 0) \;=\; \min(p_i, p_j) + 1 - \max(p_i, p_j) \;=\; 1 - |p_i -$$

Under cumulative responding, simple matching is *exactly linear* in difficulty distance: it discriminates as well between the easiest and hardest items as between neighbors. Its information horizon is the whole continuum. Real responding is stochastic rather than cumulative, which adds noise and curvature to (1), but the essential property survives across the ranges studied here: the coefficient keeps ordering pairs by difficulty distance all the way out.

The same algebra imposes an important limit on interpretation. Under complete item independence, the expected simple-matching proximity is $p_i p_j + (1 - p_i)(1 - p_j)$, which is likewise ordered by the marginals: it is largest for pairs of similar difficulty and declines as the marginals separate. An agreement-based configuration therefore orders items by difficulty *whether or not any inter-item structure exists*; the ordering is a property of the marginals, not evidence of a latent continuum. Phi does not share this behavior — under independence its expectation is zero for every pair, and no ordering is available to display. These derivations are tested directly by the permutation control of Section 5.5, and they define what a difficulty-ordered configuration can and cannot establish. Figure

1 shows the two decay profiles empirically, computed over all item pairs of the compact pool analyzed in this study: the mean simple-matching proximity declines nearly linearly across the full difficulty range, whereas the mean phi is already compressed at short distances and declines against its attainable maximum.

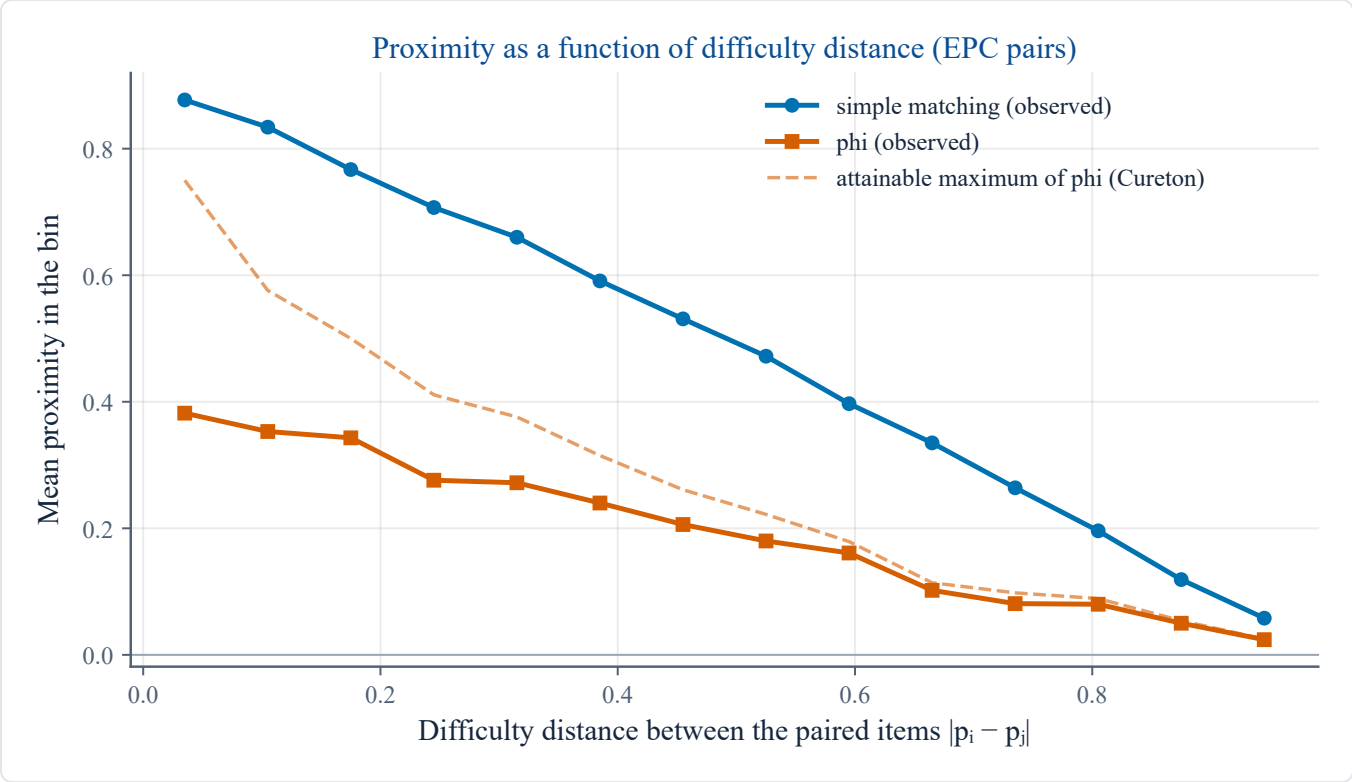

Figure 1. Mean observed proximity as a function of the difficulty distance between paired items (binned over all EPC item pairs; *N* = 1,818). Simple matching remains informative across the full range; phi is compressed under its attainable maximum (2), which itself declines with the separation of the marginals.

Phi behaves differently by construction. Its attainable maximum for a pair with marginals $p_i \leq p_j$ is

$$\phi_{ij}^{\max} = \sqrt{\frac{p_i\,(1-p_j)}{(1-p_i)\,p_j}}, \tag{2}$$

which decays toward zero as the marginals separate (Cureton, 1959). Between items far apart in difficulty, phi is compressed against a ceiling that is itself collapsing: all long-range values are squeezed into a narrow band near zero, whatever the underlying association. Phi is therefore a locally informative kernel in the sense of Diaconis et al.

(2008), and the horseshoe obtained under the phi recipe reflects that compression rather than an additional property of the items. Three predictions follow.

First, replacing phi with a full-horizon coefficient should eliminate the fold: the items should retain their order along the first dimension but form a flat band with improved fit, because the algorithm is no longer required to reproduce a large set of near-identical long-range values in the plane. Second, because the compressed long-range phi values are small quantities estimated with sampling error of order $1/\sqrt{N}$, phi-based scaling should degrade in moderate samples and improve as N grows, a prediction testable by simulation and independent of the number of items. Third, locality can be imposed explicitly on a full-horizon coefficient by declaring all proximities below a stated threshold missing, so that only local structure is fitted; this is a standard device in computational seriation, and alternating least squares scaling supports it directly through its missing-data cutoff (Takane et al., 1977). Under the locality account, this restriction should reproduce the horseshoe with the horizon as an explicit, reportable parameter.

4. Method

4.1 Item pools

Two operational pools of dichotomously scored items were analyzed (Table 1). The first is the Épreuve de Performance Cognitive (EPC; Jouve, 2005), a published 37-item cognitive test, analyzed here on $N = 1{,}818$ complete operational response records and item proportions correct spanning .001 to .981 — a compact pool with a very wide difficulty range. The second is the combined pool of the General Information Exam and the Information subtest of the Wechsler Adult Intelligence Scale (Wechsler, 2008), taken by the $N = 483$ adults of a published linking study (Jouve, 2026a): 178 items after removal of items without usable variance (154 GIE, 24 WAIS-IN), proportions correct .027 to .923, with Rasch item locations from the study's concurrent calibration spanning 7.4 logits. The two pools thus cross the design: compact versus wide, large sample versus moderate.

Table 1

Item pools and analysis design

Pool	Items	*N*	*p* range	Difficulty reference
EPC	37	1,818	.001–.981	Proportion correct
GIE + WAIS-IN	178 (154 + 24)	483	.027–.923	Rasch location, concurrent calibration

Note. EPC *N* counts complete records; GIE + WAIS-IN items are those retained by the linking study's variance filter (proportions correct in .02–.98, unanswered criterion items scored 0), and its difficulty reference is the Rasch calibration of Jouve (2026a). No item content is used or reproduced in any analysis.

4.2 Proximities, scaling, and local restriction

Two similarity coefficients were computed between all item pairs: the Pearson correlation between 0/1 scores (phi) and the simple-matching proximity (proportion of identical responses). Configurations were obtained with ALSCAL (Takane et al., 1977) at the ordinal level for phi (the classic recipe) and at both interval and ordinal levels for simple matching, in two dimensions, one matrix, unweighted Euclidean model. Local restriction used the program's missing-data cutoff: all similarities at or below a stated quantile of the off-diagonal distribution were treated as missing, so that the monotone transformation and the configuration are fitted to the surviving local proximities only. Cutoffs at the 20th through 45th percentiles were swept on both pools.

4.3 Shape and recovery metrics

Each two-dimensional solution is summarized by four quantities: Kruskal's Stress (formula 1) and RSQ as reported by the program; the Spearman correlation ρ between the first dimension (oriented easy to hard) and the pool's difficulty reference; and a *fold statistic*, the R^2 of the quadratic regression of the second dimension on the first, which is near zero for a flat band or a closed ring and grows as the configuration arches. Because item-level scaling of dichotomous data yields high absolute Stress by construction (Spence & Ogilvie, 1973), Stress is read comparatively, never against conventional cutoffs.

4.4 Simulation and rescue conditions

To separate sample-size effects from structural ones, synthetic response matrices were generated from the wide pool's own Rasch calibration — item locations from the linking study, abilities resampled with replacement from its 483 empirical estimates — at N = 483, 1,000, 2,000, 5,000, and 20,000, and scaled with both recipes. Under this generator the data are exactly unidimensional with the real difficulty spectrum, so any residual failure is attributable to the coefficient. Augmented designs (the real 483 records topped up with synthetic ones to N = 1,818 and 5,000) were run as well. Within the phi family, three published remedies were tested on the wide pool: trimming extreme-difficulty items (retaining p in .05–.95, then .10–.90) and rescaling each coefficient by its attainable maximum, phi/phi-max (2) (Cureton, 1959; Davenport & El-Sanhurry, 1991). Values for two margin-free coefficients (tetrachoric, Yule's Q) are carried over from the author's earlier eight-coefficient comparison on the same pool. All analyses are scripted end to end with fixed seeds.

5. Results

5.1 Configuration shape under three analyses of the same pool (RQ1)

Figure 2 presents the compact pool under the three analyses. The phi-ordinal recipe (panel a) produces the horseshoe: Stress = .126, RSQ = .947, difficulty recovery ρ = .989, fold R^2 = .61. Simple matching at the ordinal level (panel b) lays the same 37 items out almost flat — fold R^2 = .007 — while recovering the continuum essentially perfectly (ρ = .999) and fitting *better* (Stress = .056, RSQ = .991); the interval solution is nearly identical (Stress = .083, RSQ = .984, ρ = .9995). The fold is thus absent precisely when the coefficient does not saturate, with no loss of fit and improved difficulty recovery. Restricting the simple-matching solution to local proximities (panel c; values at or below the 20th percentile, .367, treated as missing) restores the arc at fold R^2 = .80 — stronger than under the phi recipe — with Stress = .069, RSQ = .984, and ρ = .946. Because the response data are identical across the three panels, the differences in configuration shape are attributable to the analysis; this addresses RQ1.

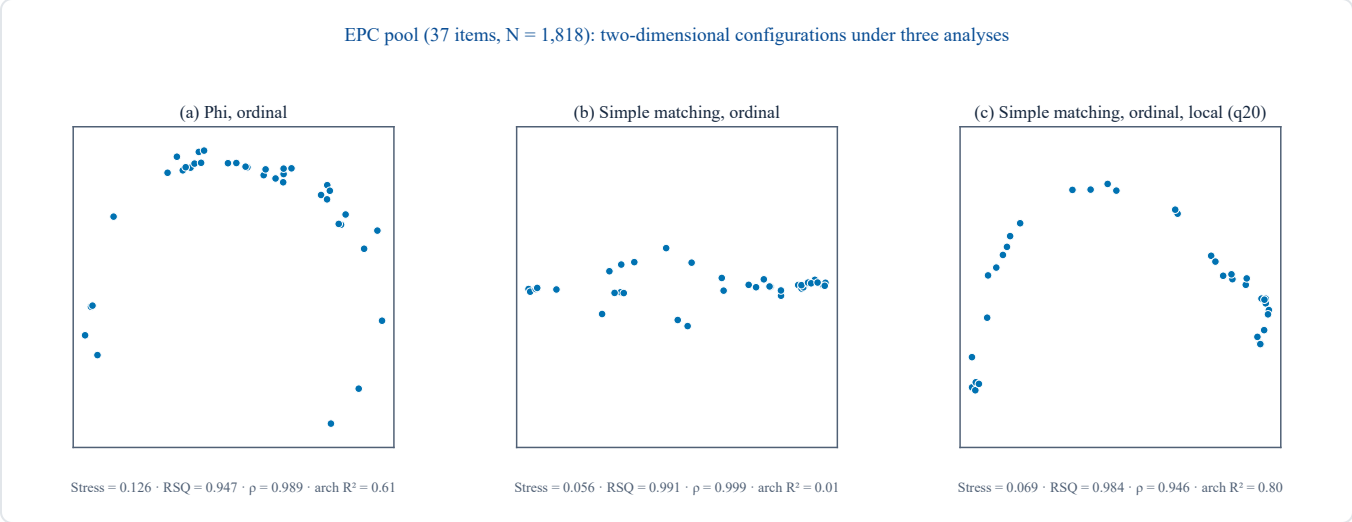

Figure 2. Two-dimensional configurations of the 37 EPC items (*N* = 1,818) under three analyses: (a) phi proximities at the ordinal level, (b) simple-matching proximities at the ordinal level, and (c) simple-matching proximities with values ≤ .367 (the lowest quintile) treated as missing. Fit, difficulty recovery, and fold statistics are given beneath each panel.

5.2 Configuration shape on the wide pool

Results for the wide pool are given in Figure 3 and Table 2. On this pool the phi recipe does not produce an interpretable configuration: Stress = .313, RSQ = .488, $\rho = .89$, no fold (panel a). Simple matching at the interval level (panel b) recovers the continuum at $\rho = .9987$ but closes the configuration into a ring; at the ordinal level (panel c), where the monotone transformation is free to stretch the long-range proximities, the ring opens into a flat band with the best fit of any full-information solution on this pool (Stress = .139, RSQ = .931, $\rho = .998$). Neither full-information solution shows a fold. With local restriction at the 25th percentile (panel d; similarity ≤ .497 treated as missing), the open horseshoe is obtained — the items ordered from easiest to hardest along the arc, with the two instruments interleaved — at Stress = .172, RSQ = .897, $\rho = .985$, fold $R^2 = .30$. The companion paper accounts for the composition of the second dimension in the full-information interval solution, where a content contrast rather than the difficulty harmonic occupies it (Jouve, 2026b); the present analyses establish the prior point that no full-information solution on this pool produces a fold.

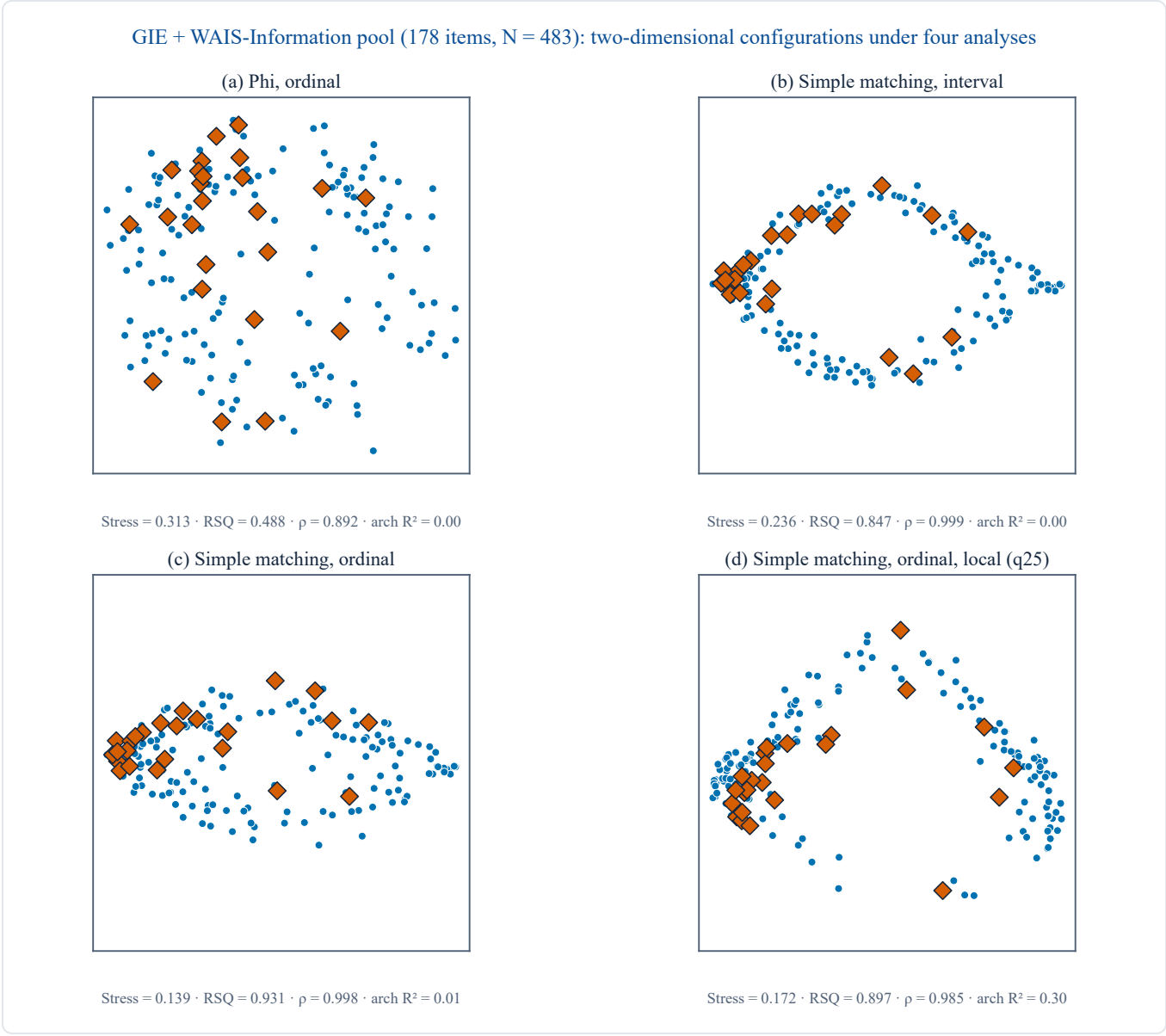

Figure 3. Two-dimensional configurations of the combined GIE + WAIS-IN pool (178 items, *N* = 483) under (a) phi-ordinal, (b) simple-matching interval, (c) simple-matching ordinal, and (d) simple-matching ordinal with values ≤ .497 (the lowest quartile) treated as missing. Diamonds are the 24 WAIS-IN items.

Table 2

Fit, difficulty recovery, and fold statistics for the principal analyses of both pools

Pool	Recipe	Stress	RSQ	ρ (dim 1, difficulty)	Fold R^2
EPC	phi, ordinal	.126	.947	.989	.61
EPC	simple matching, interval	.083	.984	.9995	.005
EPC	simple matching, ordinal	.056	.991	.999	.007

Pool	Recipe	Stress	RSQ	ρ (dim 1, difficulty)	Fold R^2
EPC	simple matching, ordinal, local (q20)	.069	.984	.946	.80
GIE + WAIS-IN	phi, ordinal	.313	.488	.892	.002
GIE + WAIS-IN	simple matching, interval	.236	.847	.9987	.000
GIE + WAIS-IN	simple matching, ordinal	.139	.931	.998	.006
GIE + WAIS-IN	simple matching, ordinal, local (q25)	.172	.897	.985	.30

Note. Stress is Kruskal's formula 1; RSQ is the squared correlation between disparities and distances; ρ is the Spearman correlation of the oriented first dimension with the pool's difficulty reference (Table 1); fold R^2 is the quadratic regression of dimension 2 on dimension 1. "Local (q*X*)" = similarities at or below the *X*th percentile treated as missing. All two-dimensional solutions, one matrix, Euclidean model.

5.3 Sample-size dependence of the phi recipe (RQ2, first part)

The Rasch simulation isolates the reason the phi recipe succeeds on the compact pool and fails on the wide one (Figure 4). With the wide pool's item locations and ability distribution, phi-ordinal at $N = 483$ performs close to its real-data values ($\rho = .941$ simulated vs. .892 observed; RSQ = .270 vs. .488) and improves monotonically with sample size: $\rho = .980$ at $N = 1{,}000$, .989 at 2,000, .996 at 5,000, and .999 at 20,000, with RSQ rising from .27 to .93. Simple matching is insensitive to sample size, reaching $\rho = .996$ at $N = 483$ and $\rho = 1.00$ at 20,000 with essentially unchanged fit. Augmenting the real 483 records with calibration-faithful synthetic records reproduces the same progression ($\rho = .987$ at a combined $N = 1{,}818$; .996 at 5,000). A complementary subsampling experiment yields the corresponding result on the compact pool: in three draws of $N = 483$ from the 1,818 records, the phi recipe remains intact twice ($\rho = .980$, .978) and fails once ($\rho = .352$). At moderate *N*, phi's compressed long-range values are too noisy to scale reliably on either pool.

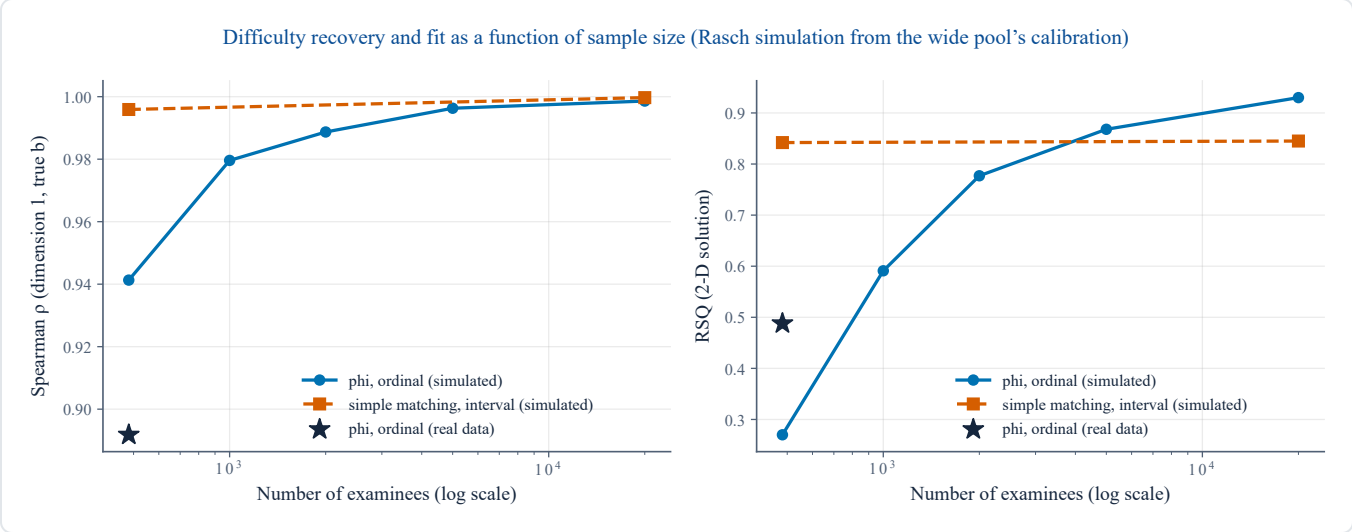

Figure 4. Difficulty recovery and fit as a function of sample size. Synthetic response matrices from the wide pool's own Rasch calibration (item locations from the linking study; abilities resampled from the 483 empirical estimates), scaled with both recipes. Left: Spearman ρ between dimension 1 and true item location. Right: RSQ of the two-dimensional solution. The star marks phi-ordinal on the real data.

5.4 Corrections within the phi family; the fold under explicit locality (RQ2, second part; RQ3)

The standard correction for phi's difficulty compression is division of each coefficient by its attainable maximum (2). On the wide pool this correction eliminates the difficulty dimension rather than restoring it: phi/phi-max scales to $\rho = .019$ (Figure 5a). The result follows from the framework: the ordering information displayed by item configurations is carried by phi's marginal dependence, and removing the dependence removes the basis of the ordering. Margin-free coefficients behave equivalently in the author's earlier comparison on the same pool (tetrachoric, $|r| = .18$; Yule's Q, .06). Trimming extreme items degrades recovery in the opposite direction — ρ falls from .892 (full pool) to .853 (items with p in .05–.95) to .685 (.10–.90) — because the compression operates along the whole continuum and trimming only shortens it. No setting of these corrections reproduces the recovery that simple matching attains at any sample size.

Explicit locality, in contrast, is controllable, with one parameter requiring disclosure. Figure 5b sweeps the missing-data cutoff on both pools. On the compact pool the fold strength peaks at the 20th percentile (fold $R^2 = .80$) and declines as more values are discarded; on the wide pool it peaks at the 25th (.30), with values of .03 to .23 at adjacent cutoffs. The resulting horseshoe preserves the difficulty ordering (dimension 1 recovers

difficulty at $\rho = .95–.99$), but its strength depends on the chosen cutoff, and the cutoff should therefore be reported whenever a locally scaled configuration is presented.

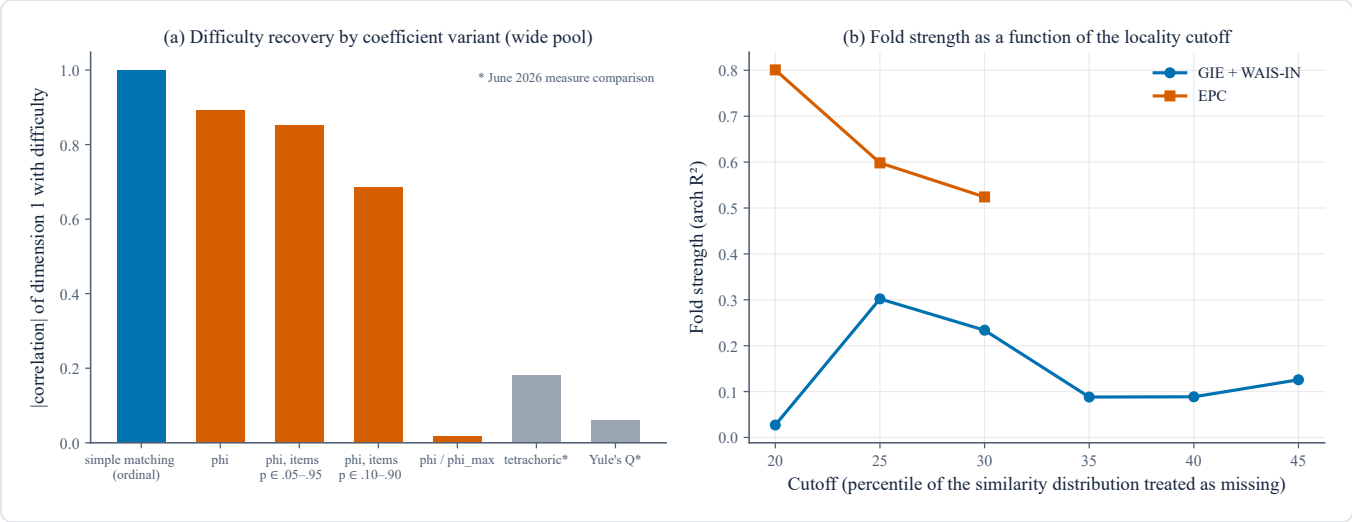

Figure 5. (a) Difficulty recovery |ρ| by coefficient variant on the wide pool; starred values are from the author's earlier eight-coefficient comparison on the same data. Margin-corrected and margin-free coefficients eliminate the continuum rather than de-biasing it. (b) Fold strength as a function of the locality cutoff (percentile of the similarity distribution treated as missing).

5.5 Permutation control: what the ordering requires

To test the boundary conditions derived in Section 3, each column of the compact pool's response matrix was permuted independently, preserving every item's proportion correct while destroying all inter-item association. On the permuted data, phi-ordinal scaling recovers nothing ($\rho = .05$, RSQ = .17): phi's configuration requires real association structure, and when the phi horseshoe appears on intact data it therefore reflects that structure. Simple-matching scaling of the same permuted data, in contrast, recovers the difficulty ordering essentially perfectly ($\rho = .996$, Stress = .075, RSQ = .985; $\rho = .964$ under local restriction at the 20th percentile) — from marginals alone, in data containing no structure whatsoever. Two consequences follow. A difficulty-ordered agreement-based configuration, with or without the fold, is not by itself evidence of a latent continuum; dimensionality must be established independently, as the linking study's factor analysis does for the wide pool (Jouve, 2026a), and as the agreement between the phi-based and agreement-based orderings does within the present compact-pool analyses. Conversely, the local-restriction procedure of Section 5.4 does not manufacture order from arbitrary

variables: applied to a pool without marginal heterogeneity or without an independently established continuum, it produces a configuration that displays the marginals, and nothing more, and it should be interpreted accordingly.

6. Discussion

RQ1. In the data examined here, the difficulty horseshoe is not a property of the item set. On both pools, the same items under a full-information coefficient produce a flat band or a ring with equal or better fit and equal or better difficulty recovery; the fold appears only when the proximity is informative locally — by the accident of phi's marginal compression, or by the analyst's explicit restriction. **RQ2.** The results indicate that the classic recipe fails on the wide pool for two compounding reasons: its long-range values are compressed into a noise-dominated band, and at N in the hundreds that noise overwhelms the scaling — a limitation that resolves only with samples in the thousands. The evidence further suggests it cannot be corrected within the phi family, because the compression and the ordering signal are carried by the same marginal dependence: corrections that remove the compression also remove the basis of the configuration. **RQ3.** The open horseshoe can be produced under explicit control — simple-matching proximities at the ordinal level, with similarities below a stated quantile treated as missing — on both pools and at both sample sizes, with the locality horizon as a reportable parameter. What the resulting configuration establishes is limited: the permutation control shows that an agreement-based difficulty ordering arises from the marginals alone, so the configuration displays an ordering but does not, by itself, demonstrate a latent continuum; that demonstration requires independent structural evidence.

This account is continuous with the classical explanations of the horseshoe rather than a rival to them. Kendall (1970) and Diaconis et al. (2008) take locality as given and derive the geometry that follows from it; the present results identify the statistical mechanism by which psychometric proximity coefficients satisfy — or fail to satisfy — the locality condition those derivations require, and show that the condition can be manipulated directly. The contrast drawn throughout is accordingly geometric, between full-

information and local-information coefficients, not an evaluative ranking of coefficients: which class is appropriate depends on whether the analyst wants the continuum displayed flat or folded, and on the sample available.

6.1 Practical recommendations

For item maps of dichotomous tests, three rules follow. First, a full-information coefficient (here, simple matching), scaled ordinally, is the appropriate default: it is robust to sample size, it recovers the difficulty ordering essentially perfectly, and its configuration — band or ring — is the conservative exhibit. Second, when the horseshoe is wanted as an interpretive display, it should be produced by explicit local restriction with the cutoff reported, in the manner of a smoothing bandwidth; an unreported horizon presents an analysis parameter as a property of the test. Third, the phi-ordinal recipe should be reserved for the conditions it requires — compact pools and samples in the thousands — and its horseshoe should be interpreted with the understanding that the curvature reflects the coefficient's compression in addition to the items' order. Fourth, a difficulty-ordered configuration produced with an agreement coefficient should be accompanied by independent evidence of dimensionality whenever it is offered in support of a continuum interpretation, because the ordering itself is available from the marginals under complete independence (Section 5.5). These recommendations do not reduce the configuration's diagnostic value: a first dimension that fails to track difficulty, or ends that do not order correctly, remains informative. They concern attribution: the fold indicates the information horizon of the analysis, the ordering indicates the marginals, and neither should be presented as more than that without corroboration.

6.2 Limitations

The evidence comes from two operational pools of dichotomous items, two proximity coefficients, and one calibrated simulation; the locality account predicts the same behavior wherever a proximity saturates along an ordered continuum, but that breadth is argued, not shown. Other coefficient families, polytomous or nominal items, and latent structures other than a dominant difficulty ordering were not examined, and the account should be treated as a hypothesis in those settings. The fold statistic is a deliberately crude summary — a quadratic fit cannot distinguish a ring from a band, which is why

the configurations themselves are shown. Alternating least squares scaling is iterative and can settle in local minima; the sweeps here used single runs per condition, and cutoff-adjacent solutions on the wide pool visibly jump between shape regimes. The permutation control is likewise a single draw. Finally, the simulation is model-true Rasch: it isolates the sample-size mechanism precisely because it removes everything else, and its synthetic records are evidence about coefficients, never about examinees.

7. Conclusion

Across two operational item pools and a calibrated simulation, the difficulty horseshoe appeared when, and only when, the proximities supplied to the scaling ceased to discriminate beyond a local horizon. Phi imposes such a horizon through its marginal compression, at the cost of substantial sampling noise at moderate *N*; simple matching imposes none and configures the continuum without a fold; explicit local restriction reinstates the horizon under analytic control. In these data the horseshoe accordingly behaved as a signature of the information horizon of the analysis rather than of the item set, and it should be produced and reported as such — with the coefficient, the measurement level, and the locality threshold stated. A permutation control fixes the interpretive boundary: agreement-based configurations order items by difficulty from the marginals alone, so neither the ordering nor the fold substitutes for independent evidence of dimensionality. Presented under those conditions, it remains what the seriation tradition established: an interpretable one-dimensional ordering displayed in two dimensions.

References

Cureton, E. E. (1959). Note on φ/φ_{max}. *Psychometrika, 24*(1), 89–91. https://doi.org/10.1007/BF02289765

Davenport, E. C., & El-Sanhurry, N. A. (1991). Phi/phimax: Review and synthesis. *Educational and Psychological Measurement, 51*(4), 821–828. https://doi.org/10.1177/001316449105100403

Diaconis, P., Goel, S., & Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics, 2*(3), 777–807. https://doi.org/10.1214/08-AOAS165

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction*. Princeton University Press.

Jouve, X. (2005). *Épreuve de Performance Cognitive (EPC).* Éditions du Centre de Psychologie Appliquée (ECPA par Pearson).

Jouve, X. (2026a). Linking the General Information Exam to WAIS Information: One crystallized-knowledge dimension and a raw-to-scaled conversion. *Cogn-IQ Research Papers.* https://pubscience.org/ps-1n0JH-be9993-znpb

Jouve, X. (2026b). Difficulty, skill, and the shape of the map: A spectral decomposition of multidimensional scaling for binary test items. *Cogn-IQ Research Papers.* https://pubscience.org/ps-1n0HX-05b978-bVCm

Kendall, D. G. (1970). A mathematical approach to seriation. *Philosophical Transactions of the Royal Society of London. Series A, 269*(1193), 125–134. https://doi.org/10.1098/rsta.1970.0091

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27. https://doi.org/10.1007/BF02289565

Spence, I., & Ogilvie, J. C. (1973). A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivariate Behavioral Research, 8*(4), 511–517. https://doi.org/10.1207/s15327906mbr0804_8

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42*(1), 7–67. https://doi.org/10.1007/BF02293745

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV).* Pearson.

Figures & tables

Table 1. *Item pools and analysis design*

Figure 1. *Mean proximity as a function of difficulty distance: the information horizon of each coefficient.*

Figure 2. *Configurations of the EPC pool under phi, simple matching, and locally restricted simple matching.*

Figure 3. *Configurations of the GIE + WAIS-IN pool under four analyses.*

Table 2. *Fit, difficulty recovery, and fold statistics for the principal analyses of both pools*

Figure 4. *Difficulty recovery and fit as a function of sample size (Rasch simulation).*

Figure 5. *Coefficient corrections on the wide pool; fold strength as a function of the locality cutoff.*

CITE THIS ARTICLE

Jouve, X. (2026). The difficulty horseshoe in item-level multidimensional scaling as a signature of local information. *Cogn-IQ Research Papers.* https://pubscience.org/ps-1n0JE-2ebae9-MqDI

Export: BibTeX · RIS · EndNote · APA · Plain text

OUTLINE

Abstract

1. Introduction
2. The Horseshoe in Theory
3. A Locality Account
4. Method
5. Results
6. Discussion
7. Conclusion

References

ARTICLE

JOURNAL
Cogn-IQ Research Papers

TYPE
Methodological Framework · Theoretical Frameworks

PUBLISHED
2026

ACCESS
Open Access

DOI
pubscience.org/ps-1n0JE-2ebae9-MqDI

TOOLS

- Full text
- Figures & tables
- Cite this article
- References

CONTINUE ON COGN-IQ

Related Educational Resources

- Multidimensional Scaling — how proximity data become spatial configurations, and how to read them.
- Item Response Theory — difficulty and discrimination as parameters of a latent continuum.